STOCHASTIC INTERACTIVE ACTIVATIOI AND THE EFFECT OF CONTEXT ON PERCEPTION

Technical Report AIP - 68

James L. McClelland

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213

The Artificial Intelligence and Psychology Project

Departments of
Computer Science and Psychology
Carnegie Mellon University

Learning Research and Development Center
University of Pittsburgh

STOCHASTIC INTERACTIVE ACTIVATION AND THE EFFECT OF CONTEXT ON PERCEPTION

Technical Report AIP - 68

James L. McClelland

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213

14 July 1989

The preparation of this article was supported by NSF Grant BNS 86-09729 and ONR Contract N00014-86-G-0146; by a contract from ONR and DARPA Number N00014-86-K-0349; by NIMH Career Development Award MH00385, and by NSF Grant BNS 88-12048. I would like to thank the members of the PDP research group at CMU for useful discussions.

This research was supported by the Computer Sciences Division, Office of Naval Research, under contract number N00014-86-K-0678. Reproduction in whole or part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; Distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AIP - 68	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Carnegie-Mellon University 6b. OFFICE SYMBOL *(If applicable)*	7a. NAME OF MONITORING ORGANIZATION Computer Sciences Division Office of Naval Research
6c. ADDRESS *(City, State, and ZIP Code)* Department of Psychology Pittsburgh, Pennsylvania 15213	7b. ADDRESS *(City, State, and ZIP Code)* 800 N. Quincy Street Arlington, Virginia 22217-5000
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Same as Monitoring Organization 8b. OFFICE SYMBOL *(If applicable)*	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-86-K-0678
8c. ADDRESS *(City, State, and ZIP Code)*	10 SOURCE OF FUNDING NUMBERS p4000ub201/7-4-86

PROGRAM ELEMENT NO	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO
N/A	N/A	N/A	N/A

11. TITLE *(Include Security Classification)*

Stochastic interactive activation and the effect of context on perception

12. PERSONAL AUTHOR(S)

McClelland, James L.

13a. TYPE OF REPORT Technical	13b. TIME COVERED FROM 86Sept15 TO 91Sept14	14. DATE OF REPORT *(Year, Month, Day)* 1989 July 14	15. PAGE COUNT 42

16. SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

Neural networks, connectionist models, perception, signal detection. (KF)

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Classically, context exerts a biasing effect on perceptual identification responses given without time pressure. Such effects are well described by classical models formulated in terms of signal detection theory or Luce's theory of choice. The classical models do not describe the actual time course of processing, however; they simply produce characterizations of asymptotic response probabilities. In this article, mathematical analysis and computer simulation methods are used to show that interactive activation models exhibit the classical effect of context when they are allowed to run to equilibrium, if there is variability in the input to the network or if there is intrinsic randomness in the processing activity of the network itself. The findings suggest that interactive activation models should not be viewed as alternatives to classical accounts, but as hypotheses about the dynamics of information processing that lead to the asymptotic behavior that the classical models describe.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT □ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT □ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION
22a NAME OF RESPONSIBLE INDIVIDUAL Dr. Alan L. Meyrowitz	22b TELEPHONE *(Include Area Code)* (202) 696-4302 22c. OFFICE SYMBOL N00014

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

Stochastic Interactive Activation and the Effect of Context on Perception

James L. McClelland
Carnegie Mellon University

The idea that perception involves the joint use of stimulus and contextual information has a long history in psychology (Pillsbury, 1897; Bagley, 1900; Miller, Heise and Lichten, 1951; Neisser, 1967) and has been incorporated in one way or another into a number of models of the cognitive-perceptual interface (Massaro, 1975; McClelland, 1987; Morton, 1969; Rumelhart, 1977).

In many experiments, the effect of context is well-described by assuming it exerts a biasing effect on perceptual identification. This was the central thrust of Morton's (1969) logogen model of word identification. Massaro, Oden, and collaborators have successfully applied a similar model called the fuzzy logical model to account for a number of findings in the perception of spoken and written stimuli (c.f., e.g., Massaro, 1979; Massaro and Cohen, 1983; Oden and Massaro, 1978).

The logogen model and the fuzzy logical model base their accounts of perceptual identification responses on signal detection theory and on Luce's theory of Choice (Luce, 1963). Both signal detection and choice theory describe asymptotic performance, rather than the time course of processing that leads to these asymptotic outcomes. In contrast, the interactive activation (IA) framework for modeling of context effects, introduced in McClelland and Rumelhart (1981) and Rumelhart and McClelland (1982) permits the formulation of a detailed account of the time course of information processing.

In the IA framework, it is assumed that perceptual processing takes place in a system of simple processing units connected by excitatory and inhibitory connections. The units represent hypotheses about the input at several levels; so for example, for speech, they might represent features, phonemes, and words. Bottom-up (feature to phoneme, phoneme to word) and top-down (word phoneme, phoneme to feature) connections allow context and stimulus information to jointly determine the outcome of the perceptual process throughout the entire network. Processing is proceeding in both directions at the same time, so that the units on different levels are continually interacting (that is, influencing each other), throughout the course of processing. The IA framework is quite similar to Grossberg's ART framework (Grossberg, 1978a), and its dynamical properties were derived from systems described by Grossberg (1978b).

The IA framework permits the development of models that can account for findings concerning both the timing and the accuracy of perceptual identification responses. Indeed, the IA framework has been applied with some success to the role of context in letter perception (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1981, 1982) and in the perception of speech (Elman and McClelland, 1986; McClelland and Elman, 1986). In both of these applications, the time course of processing plays a crucial role, either in accounting for reaction time results directly, or in accounting for results dependent on the temporal relations between presentations of context and target information.

In a recent article, Massaro (in press) has pointed out that the IA framework, as instantiated in the TRACE model of speech perception (McClelland and Elman, 1986), fails to account correctly for the quantitative form of the biasing effect of context as seen in many experiments. Massaro claims that this deficiency in TRACE is due to the assumption of interactivity inherent in the IA framework. This is an important claim since the assumption of interactivity is the very heart of the TRACE model, and its failure would undermine the entire IA framework.

In this article I analyze the relation between classical accounts of the biasing effect of context and the interactive processing mechanisms provided by the IA framework. We will see that the IA framework actually realizes the biasing effect of context that is described by classical models. The flaw in TRACE (a flaw it shares with the earlier word perception model) lies in the model's incorrect use of Luce's choice model to relate activations derived through deterministic nonlinear interactive activation processes to response probabilities. Once the model is altered so that its states are themselves probabilistic, it captures the classical biasing effect of context correctly.

In a way, this article makes a small point, introducing a correction that is required to bring the IA framework into conformity with experimental fact and descriptively adequate theory. But the fact that this correction can be made is important theoretically. The correction makes it possible to establish a clear link between the interactive activation process and classical models of perceptual identification. This link is a part of the growing understanding of the relation between the local information processing activity of each processing unit in a complex, multi-level processing system and the global behavior of the system as a whole.

In what follows I begin with a brief review of the two classical models of perceptual identification, and I describe the form that data is expected to take under these two models. While this is old ground, it is necessary background for what follows. I then demonstrate a simple model incorporating the assumptions of TRACE that fails to behave in accordance with the classical models. I then analyze this simple model and show that it can indeed conform to the classical accounts if it is properly corrected. I go on to show that the correction can be incorporated successfully into full-blown IA models such as TRACE. The discussion considers the importance of pursuing research at multiple levels of description and of attempts to understand the relations between the levels.

Classical Accounts of Context Effects

The biasing effect of context is captured by two mathematical formulations, one deriving from signal detectability theory and one deriving from Luce's theory of choice. The formulations make quantitative predictions that are virtually equivalent in a wide range of situations (Luce, 1963). The models co-exist and indeed are often discussed interchangeably (c.f., Morton, 1969; Massaro, in press). Often, basic intuitions are conveyed in terms of signal detection, but the Luce formulation is used for quantitative modeling because of its greater mathematical tractability. I will briefly review the two models as they apply to a particularly well-studied situation that will be the focus of our attention throughout, namely the effect of context on the determination of the identity of a phoneme or letter in a two-alternative forced-choice identification task (Massaro, in press). For concreteness, we will consider the case studied by Massaro, in which the alternatives are /r/ and /l/, the stimuli are a set of five phonetic segments ranging from very /l/-like to very /r/-like, and the context is either /s_i/, /p_i/, or /t_i/. The first context favors /l/ since in English there are words beginning in /sli/ but not /sri/. The third context favors /r/ since there are words beginning in /tri/ but not /tli/; and the second context is intermediate since there are words beginning in both /pli/ and /pri/.

According to the signal detectability formulation, the presentation of a phonetic segment would be seen as giving rise to a representation (e.g., some pattern of neural activity, Luce 1963) which can be placed on a continuum on which the low end represents a high (subjective) likelihood that the stimulus is /l/ rather than /r/ and the high end represents a high (subjective) likelihood that the stimulus is /r/ rather than /l/ (See Figure 1). Each stimulus condition of the experiment gives rise to a different distribution of values on the continuum. Thus in the experiment each of the five phonetic segments would give rise to a different distribution of possible values, as shown in the Figure. The distributions are assumed to be normal on the continuum, and to have equal standard deviation σ. Choice between the two alternatives occurs by determining whether the particular representation evoked on a particular trial of the experiment exceeds a criterion or cut-point on the subjective continuum. If the representation falls to the right of the cut-point, the stimulus is interpreted as an /r/; otherwise it is interpreted as an /l/. Given this formulation, the probability of the /r/ response to a particular stimulus is simply the area to the right of the cut-point under the curve representing the distribution associated with that stimulus. There is a one-to-one mapping between these areas and the distance between the mean of the distribution and the cut-point measured in standard-deviation units. This distance is just a z-score and can be determined for any obtained probability of choosing the /r/ response by consulting a standard z-score table.

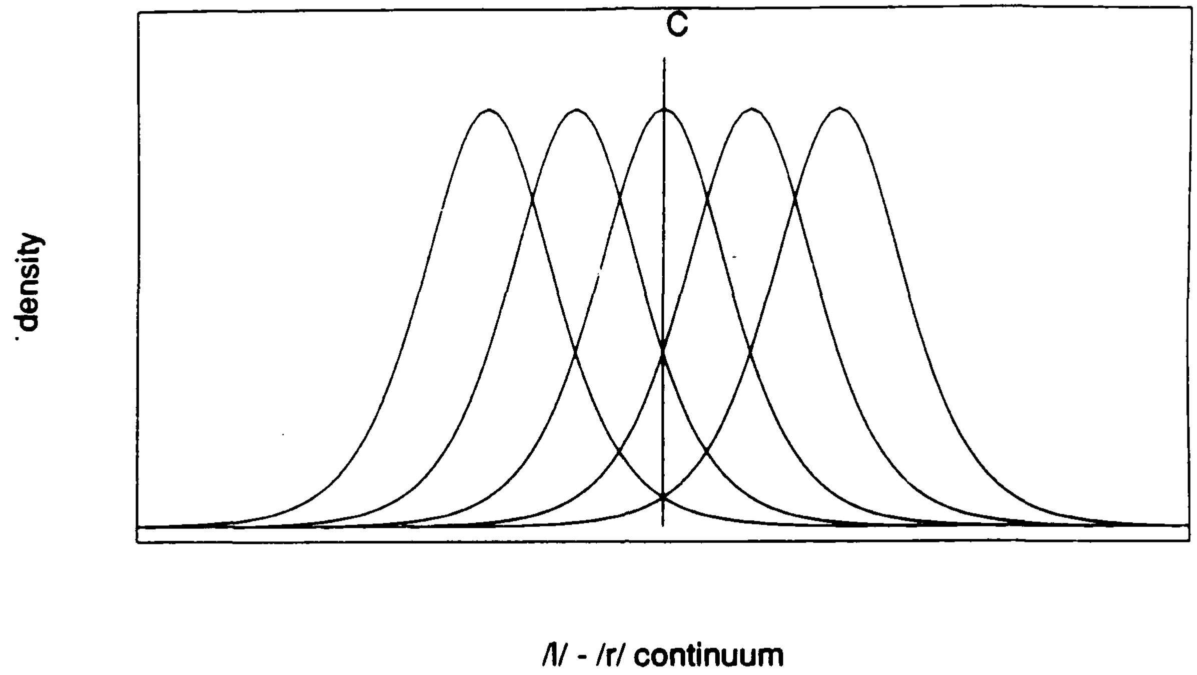

Figure 1. Distributions of relative likeness to /l/ or /r/ associated with five stimuli varying from /l/-like to /r/-like. Representations whose values on this continuum fall to the right of the cut-point C are identified as /r/, those falling to the left are labelled /l/.

The role of context in this model is assumed to be simply to shift the relationship between the cut-point and the distributions of representations produced by incoming stimuli. This can happen in either of two ways. First, the context could cause a shift in the cut-point, without altering the representations in any way. Second, the context could shift the representations themselves by a constant (opposite) amount. Obviously, these two possibilities are equivalent in the effects that they have of the areas under each curve to the right of the cut-point. Thus we get the same effect on response probabilities if we assume that the /s_i/ context shifts the representations to the right by some fixed amount, or if we assume it shifts the cut-point to the left by the same amount.

If all of this correctly characterizes perception, we should be able to fit the data we would obtain from Massaro's experiment, in which each of the five stimuli is presented repeatedly in each of the three contexts. We would need seven parameters. Five of these would represent the distances from the means of the distributions associated with each stimulus to the cut-point in one of the three contexts; the other two would represent the increment or decrement in these distances associated with each of the other two contexts. Once these parameters are set, the z-scores associated with the probability of the /r/ response in each condition would be expected to fall on three straight lines, as shown in Figure 2. The locations of the different input stimuli on the x axis would reflect the first five parameters; the separations of the three curves would reflect the other two.

We now turn to the second model. This model is based on the Luce choice theory, and has been used extensively by Massaro and his collaborators, who have called it the fuzzy logical model. In this model, no explicit assumption of variability is made. Instead, each response is assumed to have a strength, which is equal to the product of two positive terms, one associated with the stimulus, and one associated with the context. Thus, for response /r/ in our example, the

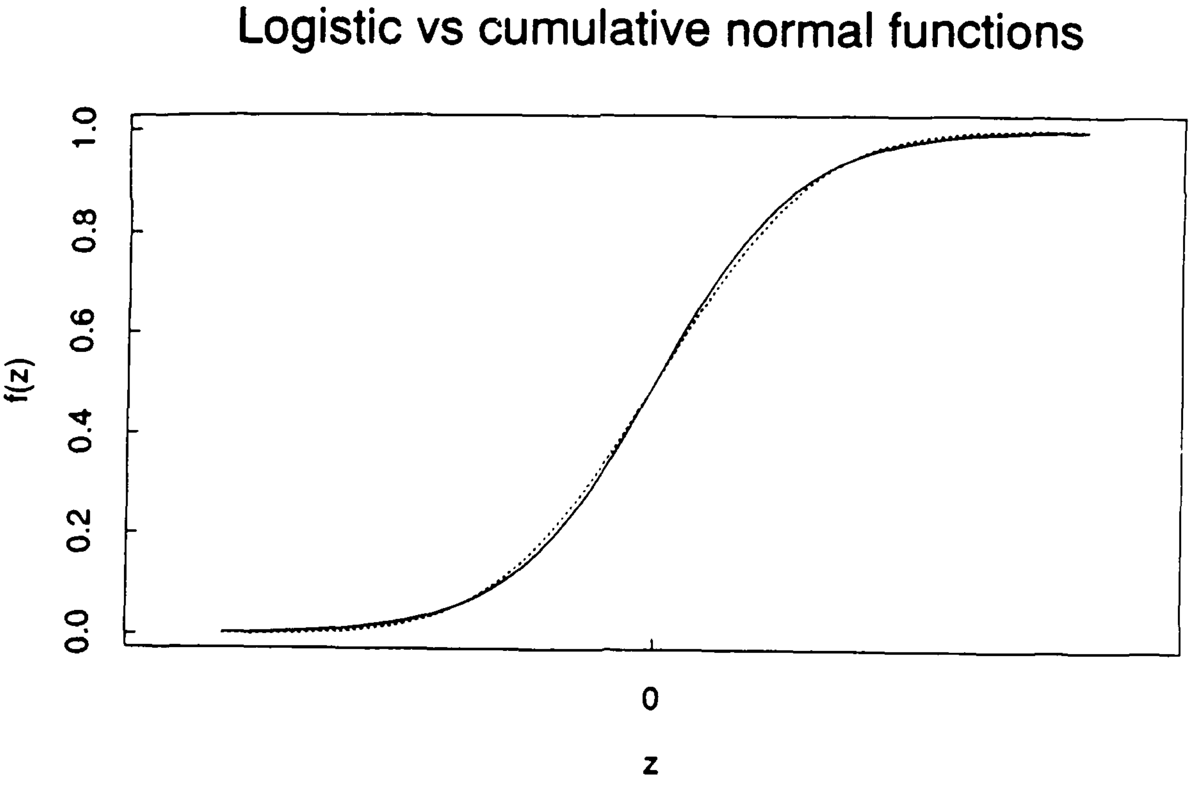

Figure 2. Expected pattern of results based on signal detection theory for neutral, /r/-biased, and /l/-biased contexts. Note that the stimulus conditions may not be evenly spaced along the x-axis, and that the biasing effects of context need not be of equal magnitude. All that is required is that the stimulus conditions can be spaced in such a way as to produce lines for the three conditions that are both straight and parallel.

strength of that response would be:

$$S_r = I_r C_r$$

and similarly for S_l. The probability of choosing response r is then given by

$$p(r) = \frac{S_r}{S_r + S_l}$$

As in the signal detection model, stimulus input is assumed to vary along a continuum from /l/-like to /r/-like. Each stimulus condition is assumed to give rise to a constant value on this continuum. For our purposes, it is convenient to represent the continuum as ranging over the positive real numbers, with 1 representing the neutral point. The points along the continuum can then be interpreted directly as representing values of I_r, with I_l being set equal to $1/I_r$, so that $I_r I_l = 1$ (it is a property of the choice equation that uniform scaling so that $I_r I_l = k$ produces the same results for any positive k).

Similarly, each context condition can be assumed to give rise to a constant value on a second continuum over the positive reals, with 1 representing again the neutral point, and with points on the continuum interpreted directly as representing values of C_r; again we let $C_l = 1/C_r$.

Now, in Luce (1963) the similarity between these two formulations is noted. To bring out this similarity, we note that we can rewrite the strength of alternative r as:

$$S_r = e^{\ln I_r} e^{\ln C_r} = e^{(\ln I_r + \ln C_r)}$$

and similarly for alternative l. Substituting into the expression for $p(r)$, dividing the numerator and denominator by S_l, and using facts about the relations between logs of products and sums of logs, it is easily shown that

$$p(r) = \frac{e^{\ln(I_r/I_l) + \ln(C_r/C_l)}}{e^{\ln(I_r/I_l) + \ln(C_r/C_l)} + 1}$$

This expression is the *logistic* function of the quantity

$$\ln(I_r/I_l) + \ln(C_r/C_l)$$

which is the sum of a stimulus term and a context term. It's form is very similar to the form of the relation between the area to the right of a cut-point under a normal curve whose mean has a distance from the cut-point equal to the sum of a stimulus term and a context term. The similarity of the two distributions is illustrated in Figure 3.

In essence, each of the two classical models assume an essentially independent influence of stimulus and context information. In the signal detection formulation, this independence exhibits itself as an additive effect on the representation of the stimulus on a continuum that represents the relative likelihood of one alternative compared to the other. In the choice model, the independence is captured by the assumption that the response strength of an alternative is the

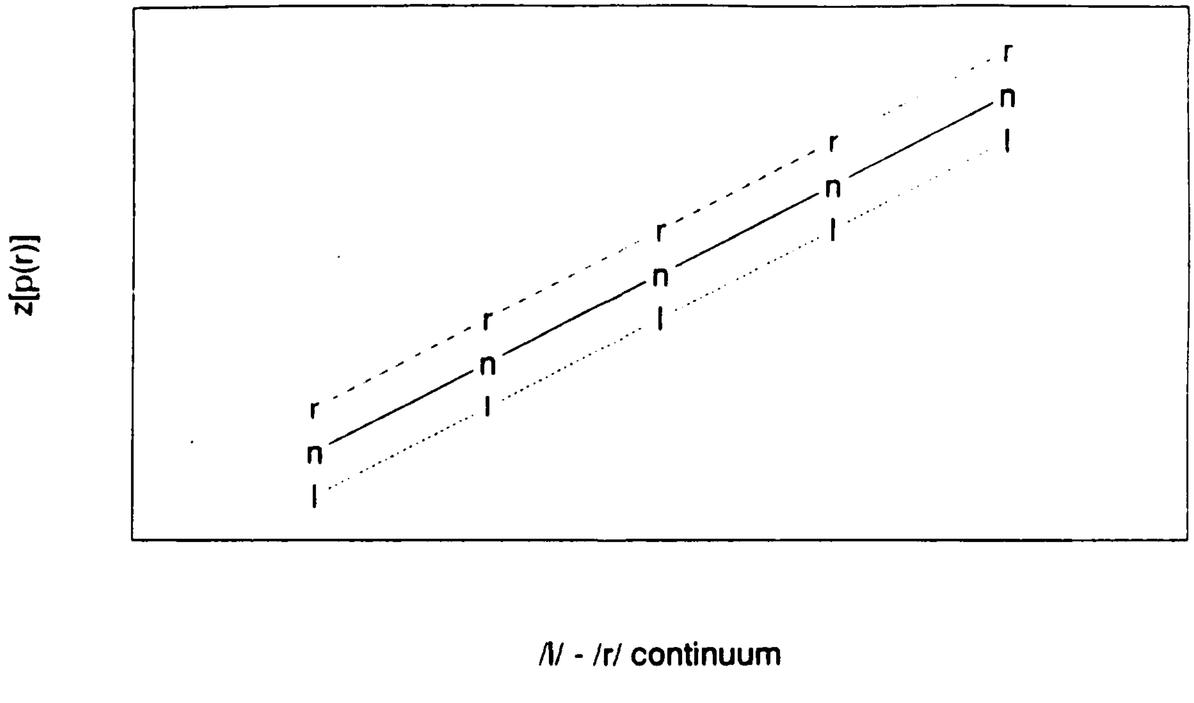

Figure 3. A comparison of the logistic function of a variable z/v (solid curve) and the area under the normal curve to the right of z (dotted curve). For both the signal detection model and the choice model, z is equal to the sum of the contextual and stimulus influences. The scale factor v is chosen so that the curves are as similar as possible.

product of two terms, one for the context and one for the stimulus. Both also assume a transformation that carries points on this continuum into response probabilities. In the signal detection formulation, points on the continuum are deterministically mapped to choices; variability arises in the representation of the combined influence of stimulus plus context. In the choice formulation, points on the continuum are deterministically related to the product of the stimulus and context terms, but response probabilities are then subject to variability arising from the use of a probabilistic decision rule.

In discussing the two models, Luce notes that both only model asymptotic choice performance. In other words, they do not provide any characterization of the time course of information processing, but only the outcome. The search for a model that has the same asymptotic behavior but which also provides a characterization of the dynamics of processing would seem in this light to be worthwhile.

Context Effects in IA Models

The IA framework is certainly one that provides a characterization of the dynamics of information processing. However, as Massaro (in press) points out, this framework, as instantiated in the letter perception model and in the TRACE model, does not produce the additive effects of context and stimulus information illustrated in Figure 2. Rather, the model systematically enhances differences at the boundary between the two alternatives while diminishing differences at the extremes. This interacts with the effects of context, which shifts the point along the /l/-/r/ continuum at which the boundary between the two alternatives falls. The effect is illustrated in Figure 4. The results shown in Figure 4 were obtained not with the actual TRACE model but with the simple network shown in Figure 5. In this network, there are two sets of phoneme level units and one set of word-level units. The first set of phoneme level units contains detectors for /s/, /p/, and /t/ while the second set contains detectors for /l/ and /r/. The first set will be called the context units and the second set the target units, since in the simulations the task will be to determine whether the "word" presented ends in /l/ or /r/. At the word level, there are detectors for the "words" /sl/, /pl/, /pr/, and /tr/. Thus /p/ serves as a neutral context equally consistent with /r/ and /l/, whereas /s/ favors /l/ and /t/ favors /r/. There are bi-directional, excitatory connections between units that are mutually consistent (so that /s/ and /l/ are each connected to the /sl/ unit, etc.). At the word level, all the units are mutually inhibitory; at the phoneme level, all the units are mutually inhibitory within each set.

This network is highly simplified compared to the full TRACE model. We study the simplified case first since it is sufficient for exploring the problems with the existing formulation of TRACE and for examining how those problems may be solved. At the end of the paper, I will return to the full TRACE model and show that when the repair that fixes the small model is applied to the full model, it fixes the full model as well.

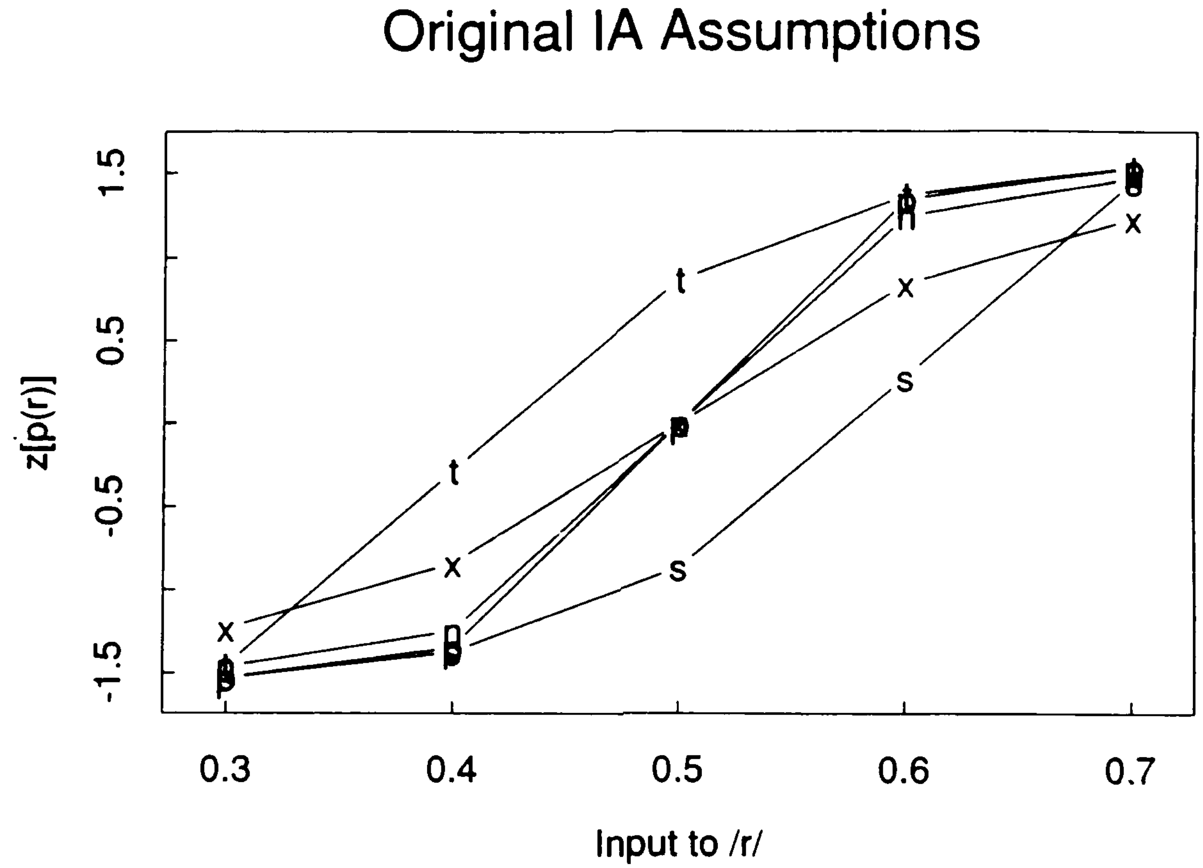

Figure 4. The joint effects of context and stimulus information, from the simple IA network shown in Figure 5. Similar results are obtained when using the full TRACE model (Massaro, in press). The graph shows the z-transformed probability of choosing the /r/ response, for each combination of the stimulus and context conditions. The curves labeled s, p, and t refer to the /s/, /p/, and /t/ context conditions respectively. The curve labeled n refers to the no context condition and the curve labeled x refers to the condition in which the connections between the target phoneme units and the word units are removed.

The network is implemented using the **iac** program of McClelland and Rumelhart (1988). This program embodies the same processing assumptions as the TRACE model and the interactive activation model of word perception. These assumptions are closely related to networks studied by Grossberg (e.g., Grossberg, 1978a, 1978b). The **iac** program was augmented to include the mechanism that translates activations into overt responses that was used in both TRACE and in the letter perception model.

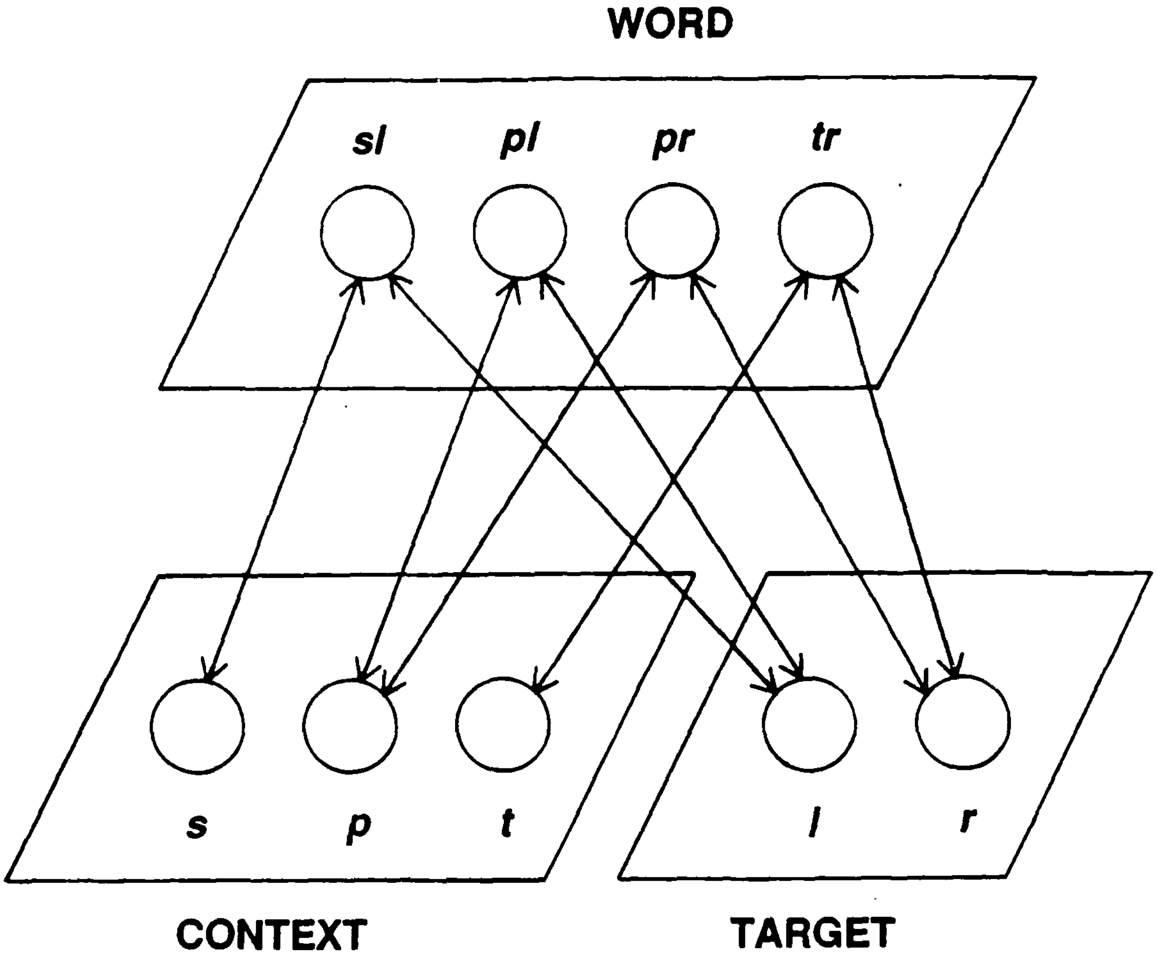

Figure 5. The simple network used throughout this paper for studying the joint effects of stimulus and contextual information in IA networks.

Model Details

In the TRACE model, inputs are actually presented as though they are arising from a time varying acoustic signal that is spread out in time. In the present, simplified situation, inputs are turned on all at once and left on until activations reach asymptote. This is more nearly equivalent to the experimental paradigm used by Massaro (1979), in which the targets were visual pattems that varied between two altemative letters, displayed together with contexts consisting of other letters.

Processing occurs as follows. Before each stimulus presentation, activations of all units in the network are set equal to their resting activation value, and extemal inputs are presented to selected units for processing. Processing then begins. Processing occurs through a sequence of time steps. At each time step, each unit computes its net input from other units based on their activation at the end of each time step. The net input to unit i is:

$$net_i = \sum_i w_{ij} o_j + ext_i$$

where w_{ij} is the connection weight from unit i to unit j, o_j is the greater of 0 and the activation of unit j, and ext_i is the external input to unit i. The connection weights are +1 for excitatory connections and -1 for inhibitory connections.

Once the net input to all units has been computed, activations are updated as follows:

If (net sub i > 0):

$$\Delta a_i = I(M - a_i) net_i - D(a_i - r); \qquad (1)$$

Otherwise,

$$\Delta a_i = I(a_i - m) net_i - D(a_i - r).$$

Here M refers to the maximum activation rate, m refers to the minimum activation rate, r refers to the resting activation level, and I and D are constants that scale the relative size of the influences of the inputs to units and of the tendency to decay back to rest respectively. (The values used for these parameters are generic: M=1; m=–.2; r=–.1; I=.1 and D=.1).[1]

For the simulations under study here, it is assumed that the subject is choosing which of the two phonemes /l/ and /r/ occurred as the second phoneme in the word. The instantaneous probability of choosing each response is calculated at each time step using the following formulae:

$$p(l) = \frac{e^{k \bar{a}_l}}{e^{k \bar{a}_l} + e^{k \bar{a}_r}}, \quad p(r) = \frac{e^{k \bar{a}_r}}{e^{k \bar{a}_l} + e^{k \bar{a}_r}} \qquad (2)$$

Here $\bar{a}_l$ and $\bar{a}_r$ refer to running averages of the activations of the phoneme units representing the alternatives l and r respectively. The running average for each unit is set to the resting activation

[1] The program has separate parameters for the excitatory, inhibitory and external input: all three were set to 0.1.

level at the beginning of each simulation run and is updated as follows after updating the activations of units in each time step:

$$\bar{a}_i(t) = \lambda a_i(t) - (1-\lambda)\bar{a}_i(t-1)$$

The value of λ was set to 0.05.

The results shown in Figure 4 are based on 25 simulation runs, factorially combining 5 input conditions with 5 context conditions. The five input conditions involve different choices of input to the units /l/ and /r/. The values for /r/ were .3, .4, .5, .6 and .7, with the value for /l/ set equal to 1 minus the corresponding value for /r/. Thus in the first input condition, the input favored /l/ while in the fifth it favored /r/; the middle condition is exactly balanced.

The five context conditions consist of one that biases processing in favor of alternative /l/, one that biases it in favor of alternative /r/, and three that are neutral. The biased contexts are those in which either context unit /s/ or context unit /t/ receives external input. The unbiased contexts are: a) one in which the context unit for /p/ receives external input; b) one in which no context unit receives external input; and c) one in which the connections between the phoneme and word levels have been severed so there is no interaction between the two levels and thus no possibility of any contextual input to the /r/ unit or the /l/ unit. The first three context conditions are similar to those used in Massaro's experiment; the others are added to aid in understanding the processes that are occurring in the network.

It is apparent in Figure 4 that the IA network does not produce a set of parallel lines relating the z-transformed probabilities to the stimulus conditions, as would be expected if it conformed to the classical models. Even the baseline case where there is no interaction with context at all (curve labeled x) is distorted. Since the inputs are evenly spaced, the z-scores of the associated response probabilities should fall on a straight line, but they do not. There is a steep transition across the midpoint on the continuum, and a leveling off at the extremes, even here. The presence of context that is neutral with respect to the two alternatives (p) or even the mere presence of mutual connections to the word level in the absence of any contextual input (n) whatsoever produces further distortions in the response probabilities compared to the baseline case. Context favoring /r/ shifts the distorted curve to the right (curve labeled t) and context favoring /l/ shifts the curve to the left.

Activations and Response Probabilities in IA Networks

What is the cause for these systematic discrepancies from the results? In the interactive activation model and in TRACE, the activation process itself is deterministic. Of course, this assumption is unrealistic. In fact, one can view this assumption as an approximation that allows us to examine, in a single simulation run, a measure of the central tendency of an ensemble of noisy activation processes. The question then arises, how should these measures of central tendency be related to observed choice probabilities? A natural assumption is that subjects choose the most active alternative on each trial. If this was in fact what happened, random variability would have the effect of introducing probabilistic responding.

Rather than deal with this variability directly, TRACE and the word perception model treated the deterministic activations in the network as inputs to a probabilistic readout process that translates activations of units into response probabilities, according to Equation 2. Probabilistic readout applied to deterministic activations can closely approximate the results of choosing the most active alternative in the presence of noise under some conditions. But it does not do so when inhibitory interactions between units and non-linear activation assumptions are added. This is where the TRACE model went astray.

I shall establish this fact by first showing that the biasing effect of context is captured correctly by our simple network when variability is introduced into the input to the network. I will then consider exactly what was wrong with the earlier formulation. Later sections of the paper will extend the results by considering variability intrinsic to the processing activity of the network.

Input Variability

The first case we will consider is the case in which the variability is in the input to the network. The model is modified as follows. On each trial, it is assumed that the input value along the input continuum is perturbed randomly by an amount that is normally distributed with standard deviation σ. The perturbed input is then applied to the network, and the activation process is allowed to proceed as before. After a number of time steps, the unit corresponding to the choice alternative with the largest running average is chosen as the network's response. With these modifications we must repeat each condition of the simulation experiment described previously many times to determine the probability of choosing each response in each condition.

Simulation. The simulation employed a 4 context by 5 input-value experimental design like the one described above but excluding the condition in which the phoneme units are isolated from the rest of the network.[2] The procedure was modified as follows. On each trial, the external inputs to the target units were shifted along the input continuum by a normally distributed random amount with standard deviation σ=.141421.[3] The simulation was allowed to run 60

[2] Simulations not reported verify that this excluded condition produces results equivalent to these in the other two neutral conditions.

[3] This value (which is just .1 times the square root of 2) was chosen because it produces percentages of choices of

cycles on each trial. At that time, the most strongly activated target unit was chosen as the network's response. 10,000 trials were run in each of the 20 combinations of context and input conditions, for a total of 200,000 simulated trials altogether (this takes overnight on a Sun 3/60).

The results of this simulation, shown in Figure 6, conform exactly to the expected form based on the classical models. The two neutral context conditions superimpose on each other, and each falls within the 95% confidence interval of the expected pattern of results, which is indicated by the straight lines on which the points superimpose. The two unbiased context conditions have been fit by a single straight line through the point $z[p(r)]=0$; the slope of this line is derived from the signal detectability analysis given below. The degree of shift up or down for the curves for the two biased contexts is fit to the simulation results, but due to the symmetry of the network the shift up for the /r/-biased context is constrained to be equal to the snift down for the /l/-biased context. So there is only one parameter estimated in fitting the simulation results to the predictions of classical models. The fit accounts for 99.97% of the variance in the z-transformed response probabilities.

Analysis. To understand what is happening in this situation, let us begin by looking at what happens if we isolate the /r/ and /l/ units from the other units in our simple net, and then carry out a series of simulation trials. On each trial, we present an external input ext_r to the /r/ unit and $ext_l = 1-ext_r$ to the /l/ unit; we let the network settle for 60 trials, and we choose as our response the alternative with the largest running average activation at this point. Across the series of trials, let us vary the input in small steps, from values strongly favoring /l/ (ext_r=0, ext_l=1), to values strongly favoring /r/ (ext_l=0, ext_r=1) (See Figure 7). Since processing is deterministic, a given point on the input continuum always produces the same result: either the /l/ unit will be the most active or the /r/ unit will be. At the bottom of the continuum the /l/ unit will be more active than the /r/ unit, but as we move along the continuum, we will gradually increase the input to /l/ and reduce the input to /r/. Eventually we will reach the point where ext_r=ext_l. For the case in which the /r/ and /l/ units are isolated from the rest of the network, the /l/ unit will be the most active for all cases in which ext_r is below the neutral point and the /r/ unit will win for all cases in which ext_r falls above it. Thus the network partitions the continuum. The cut-point in this case falls at the point where $ext_r = ext_l = .5$. This is indicated by the vertical line labeled N in Figure 7.

Now, let us reinstate the connection between the /r/ and /l/ units and the rest of the network, and introduce a fixed context input, corresponding to an external input of 1.0 to one of the three context units, or a null context of 0.0 input to all three units. In our little network, the inputs and the parameters of the net have been chosen so that, although the contextual input can influence performance, this influence is not so strong that it prevents the /l/ alternative from winning when

the /r/ alternative ranging from about .02 to .98; outside this range the probability of one of the two alternatives becomes too small to sample reliably, and small changes in probability produce very large distortions in z-scores.

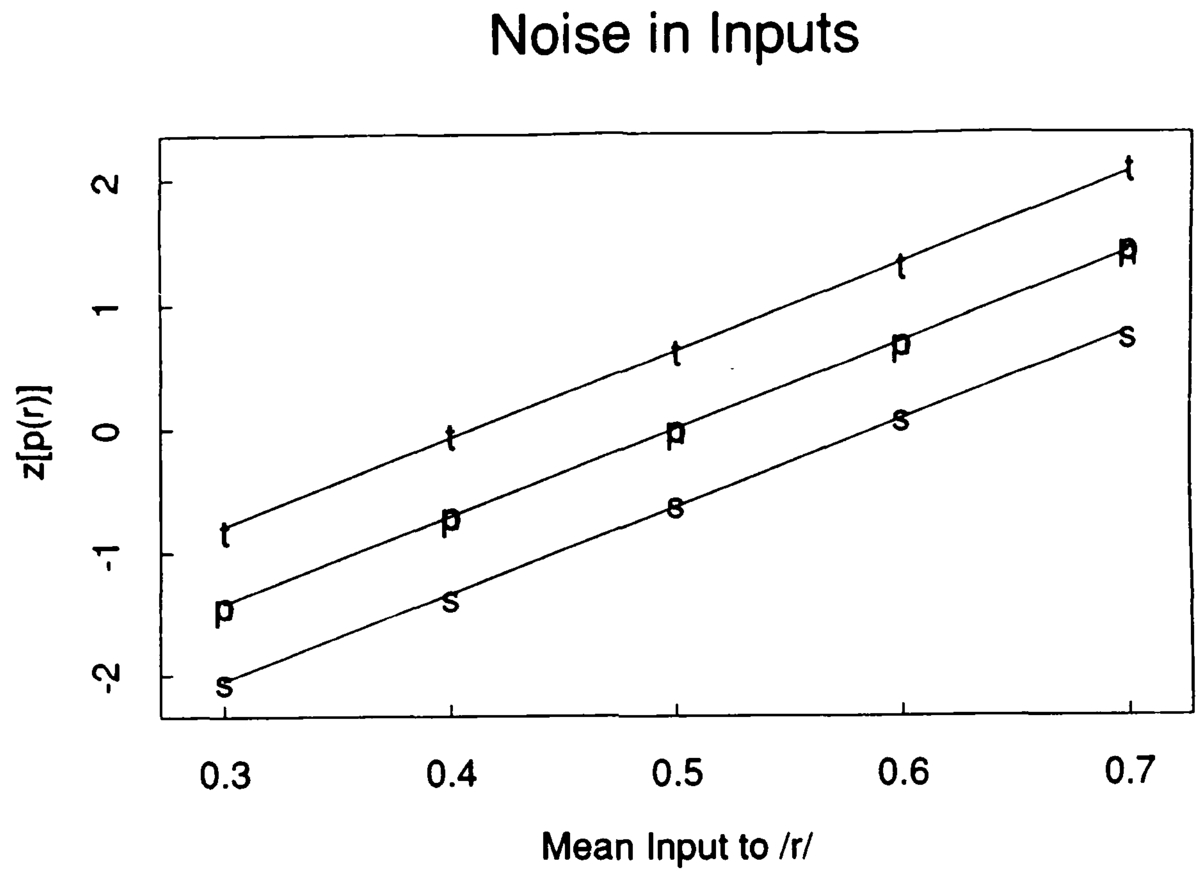

Figure 6. The z-transformed probability of choosing the /r/ response, for each context-by-stimulus combination in the simple IA network. The source of variability is the external input to the model, which is perturbed around a mean value on each trial. Labels on curves are as described in the caption to Fig. 4.

ext_r=0 and ext_l=1, or the /r/ alternative from winning when ext_r=1 and ext_l=0. Now, assume that we repeat the series of simulation runs described above again. On each trial, we present the same fixed context, together with an input on the /l/-/r/ continuum. Once again, each input will give rise to a particular final pattern of activation, in which one of the alternatives has a stronger activation than the other. The activation of the /l/ unit decreases as ext_r increases, and the activation of the /r/ unit increases. There is again a point which we will call *B* in the input continuum that divides the cases in which the asymptotic activations in the network favor /l/ from the cases favoring /r/. In Figure 7, a case in which the context favors /r/ is illustrated. The point here is not to consider just how much or even in what direction the interactions with the rest of the network will shift this point, but just to note that each context does have the effect of picking a point along the input continuum such that inputs to the left of that point favor /l/ and inputs to the right of it favor /r/.

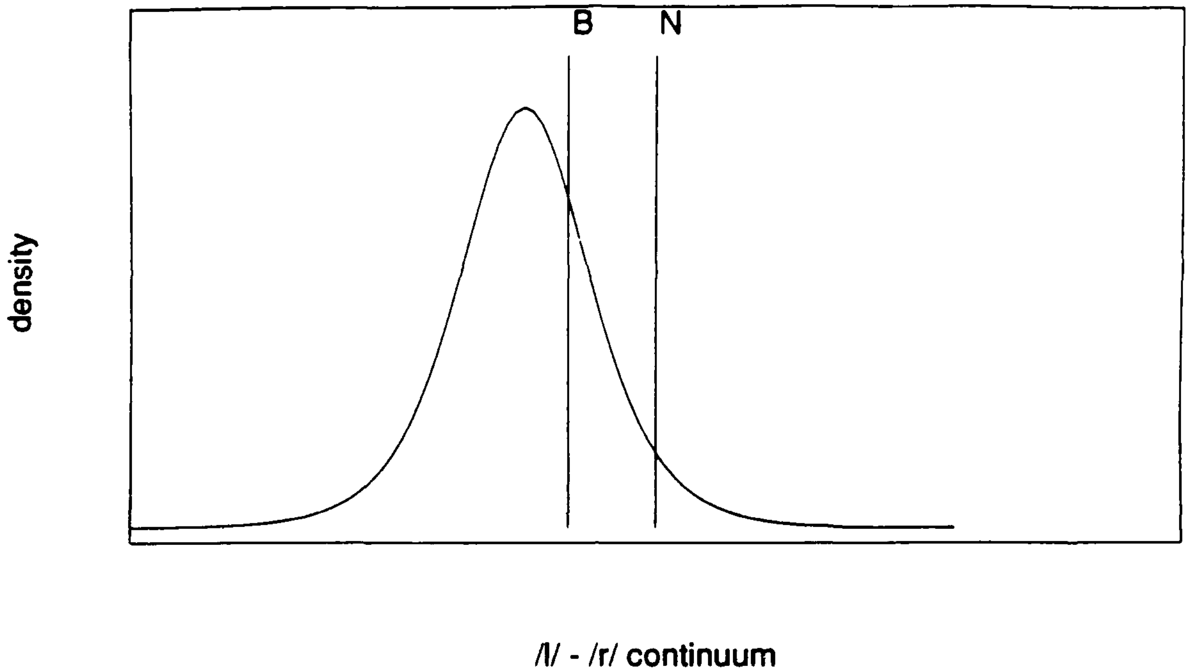

Figure 7. The input continuum along which values of ext_r are chosen. In the illustrated stimulus condition, ext_r is chosen from the distribution shown with mean μ_r=.35 and standard deviation σ=.1. N represents the cut-point in the absence of any context and B represents the cut-point in some biasing context, which in this instance happens to favor /r/, thereby shifting the cut-point to the left on the stimulus continuum so that more inputs are treated as /r/.

Now, we turn to an examination of the effect of variability in the input, given a particular context that splits the continuum at point B. Consider a value ext_r chosen from a normal distribution with mean μ_r and standard deviation σ. As before ext_l is just equal to $1-ext_r$. Assume that each stimulus condition gives rise to a different value of μ_r. Then the probability that ext_r will exceed B is just the area to the right of the cut-point B under the normal curve with mean μ_r and standard deviation σ. In sum, the context effectively shifts the criterion for choosing the /r/ alternative by the amount $(B-N)/\sigma$, and the effect of this on response probability is exactly the effect ascribed to context in signal detection theory.

This analysis allows us to specify exactly how much a change in the mean input μ_r will influence performance, for a given value of σ. Specifically, we can calculate how much a change in the μ_r will alter the z-transform of response probability: The change is simply the size of the change in the value of μ_r, divided by σ, the standard deviation of the noise. This calculation gives us the slopes of the theoretical curves relating $z[p(r)]$ to values of μ_r in Figure 6.

The analysis given above is based on the observation (tested through simulation) that for all four of the contexts (as well as for the case in which the units for /r/ and /l/ are isolated), the following conditions hold: (a) the /l/ unit is the most active unit when ext_r equals the lower bound of 0; (b) the /r/ unit is the most active when ext_r equals the upper bound of 1; and (c) the activation of the /l/ unit decreases monotonically and the activation of the /r/ unit increases monotonically as ext_r is increased from 0 to 1. The analysis also depends on the assumption that the distributions of actual inputs are normally distributed with equal variance in all conditions.

The argument will extend to any deterministic activation network where the units representing the alternatives receive direct external input perturbed by normally distributed noise and where the parameters and architecture are such that the network adheres to these three conditions for some upper and lower bound on the values on the continuum along which input varies. It is not easy to specify the exact conditions that are required for conditions (a)-(c) to hold. They hold for the network under consideration here, but in general they will depend on parameters and other details. For networks in which they hold though, we have the strong result that context will exert the classical biasing effect when variability is due to noise in the input to the network.

Where did the original formulation of the IA framework go wrong?

In the original formulation of the IA model (McClelland and Rumelhart, 1981), we treated running average activations of units as equivalent to the log of response strength, in the sense of Luce (1963). These activations are not, however, equivalent to the logs of Luce response strengths, because they are not simple sums of stimulus and context effects. The non-linear activation and competition processes in the interactive activation model distort this correspondence.

To illustrate this problem, we first consider a very simple network consisting of 2 units, one a detector for /r/ and the other for /l/ as in our example. Imagine that the units are simplified so that their activations are simply equal to their net inputs, and imagine that there are no connections between them or to any other units. Suppose that there are two external inputs to each unit, let us call them r_1 and r_2, l_1 and l_2, and finally assume that we want to determine which unit has the largest input. The following two methods of introducing probabilistic performance with respect to this network yield equivalent results:

(a) Assume that, in addition to the two inputs, the /r/ unit also receives a random perturbation, distributed normally with mean 0 and standard deviation σ, and the /l/ unit receives an equal and opposite perturbation. Because of the perturbation, the most active unit may not be the one with the largest (pre-perturbation) input. We run many such trials and compute the probability that the /r/ unit has the strongest activation.

(b) Assume that there is no perturbation, and the activation of each unit is simply set to the sum of its inputs. We then exponentiate the activation of each unit, and treat these exponentiated activations as strengths in the Luce sense. We can then calculate the probability that we will choose /r/ according to the choice equation:

$$p(r) = \frac{e^{ka_r}}{e^{ka_r} + e^{ka_l}}$$

These the two variants of the simple two-unit network exhibit the same correspondence between the signal detection and choice models that was observed earlier. That is, for any choice of perturbation size σ, there is a value of k that produces indistinguishable results. Because of this, it appears that one could use a formulation of type (b) to calculate a close approximation to the expected outcome of process of type (a).

However this correspondence no longer holds if we alter the situation and add a few characteristics of IA networks: We insert inhibitory connections between the two units, start each trial with the activations of the units at rest, and then update the activations gradually according to the nonlinear activation rule given in Equation 1, so that the activation of each unit is driven up or down as a function of its net input and of the current activation level.

These changes have no effect on choice probabilities under (a) but distort choice probabilities under (b). Under (a) the unit with the strongest (perturbed) input is still going to be the most active. The inhibitory connections serve to amplify small differences in the activations of the units, and the nonlinear processing assumptions keep extreme values bounded; in short, the mutually inhibitory interactions among units with bounded activations produce what is effectively a choice of the unit with the strongest input, as many researchers have noted (Feldman and Ballard, 1982; Grossberg, 1978b).

What happens under (b), however, is that the Luce strength of each response is no longer simply equal to the exponential of the sum of the contribution of the two cues. The non-linear activation process distorts this correspondence, so that the activation is no longer just the sum of the inputs. As the input to a unit gets larger, its activation begins to level off. This is the reason for the flattening of the curves in Figure 4 at extreme values. The sharp transition across the middle of the figure is due to the mutual inhibition, which accentuates small differences in inputs.

It is worth taking note of the fact that, in this example, these distortions occur even in the absence of any interactive processing whatsoever. Interactive processing does accentuate these distortions, as Figure 4 illustrates, but they are present even without interactivity and they only occur if the Luce choice model is applied to the output of the interactive activation process.

Discussion. To summarize, an IA network that has varying external inputs and selects the most active unit from units representing the available response alternatives acts as a signal detection mechanism in which contextual inputs act as biases. The mechanism performs the selection of one of the two alternatives by accentuating differences in their activation. Generation of an overt response amounts simply to picking out the most active unit from among those representing the alternatives.

One might wonder why there is any point in including the non-linear processing and the mutual inhibition. The reason for the non-linearity is that, when there are bi-directional connections, positive feedback can cause runaway activations that grow without bound. The non-linearity of Equation 1 prevents this -- without distorting the signal detection characteristics of the network. The inhibition is necessary to maintain differences in activation among units whose activations are bounded; without inhibition, activations in an interactive activation network generally spread until everything is maximally activated.

Accepting that the dynamical assumptions of IA networks are necessary, it may still be somewhat counter-intuitive that an interactive model can produce the classical context effects. The interactivity does, after all, amplify differences in the inputs to units, yet from the point of view of a signal detection analysis, there is no such amplification, only a criterion shifting effect. What is misleading about this way of thinking is that when there is variability in the input to the net this amplification applies to the whole perturbed signal, and does not separate signal from noise. The amplification accentuates differences in activation between the alternatives, but does not determine which alternative will come out ahead.

The failure of the TRACE model and the letter perception model as initially formulated to capture the biasing effect of context was due to their use of Luce's choice model on variables to which it is no longer applicable; these variables -- the activations of units that result from the activation-competition process that is present in IA networks -- *already reflect* the operation of the selection mechanism inherent in the network. The Luce choice model can describe the probability that a particular choice will be made given the set of inputs to each unit, but it is correctly applied only before the non-linear activation and competition process, not after it.

The case that interactive activation mechanisms implement the process of selecting among alternatives that is described by the classical models has been developed thus far for a limited case in which the noise is in the input to the network. Processing inside the network is strictly deterministic. The next section considers the effects of noise arising from the processing activity itself.

Stochastic Interactive Activation

In signal detection experiments, where the experimenter actually presents a faint tone, let us say, against a background of white noise, it may be plausible to view the variability as lying outside of the observer, in the stimulus itself. But in experiments like the one reported by Massaro (in press), the stimuli themselves do not really vary from trial to trial, or at least they do not vary much; probabilistic performance presumably arises largely because of intrinsic variability in the perceptual mechanisms. It seems important, then, to see how interactive activation mechanisms fair in the face of intrinsic noise.

To get an initial look at this matter, we need only modify the simple network again, as follows. We simply assume that at each time step, the net input to each unit has an additional term, consisting of a small amount of normally distributed random noise:

$$net_i = \sum_i w_{ij} o_j + ext_i + N(0, \sigma)$$

I will call IA networks with this property *stochastic interactive activation* networks.

Simulation. The simulation repeated the 5-input by 4-context design described above, under the following conditions: On each trial, the external inputs to the target units were supplied as in the deterministic case. The simulation was allowed to run for 60 cycles on each trial; at the end of this period activations still vary from time step to time step, but they have reached an equilibrium in which the distribution of possible states has stabilized (this fact was ascertained empirically). At this point, the unit with the largest running average activation is chosen as the network's response. As before, 10,000 trials are run in each context by stimulus input condition. The value of σ used was .14.

The results of the simulation are shown in Figure 8. We can see that the network continues to adhere to the classical pattern, even though the noise is now intrinsic rather than extrinsic to the processing activity. The straight lines fitted to the simulation results account for 99.97% of the variance.

Analysis.[4] I know of no technique for proving that the IA network used in the simulation just presented actually must conform to the classical models, but it is easy to show that a variant of the same network actually corresponds to the Luce choice theory. In this variant, we turn the network into a *Boltzmann machine* (Hinton and Sejnowski, 1983, 1986). In the Boltzmann

[4] I would like to thank Terry Sejnowski for pointing out the relevance of Boltzmann machines to the present topic. Hinton (personal communication) established some related findings but they were never published.

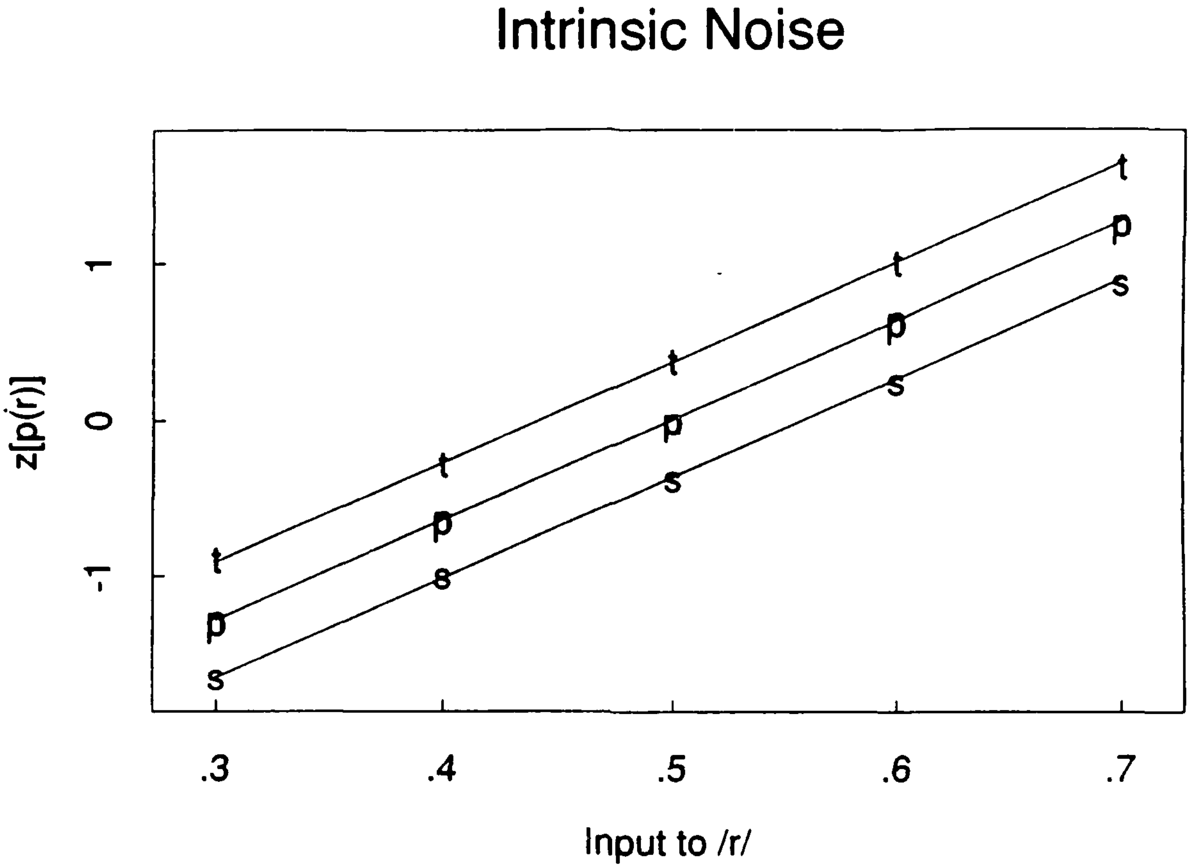

Figure 8. The z-transform of the asymptotic probability of choosing the /r/ response, for each stimulus by context combination, under conditions of intrinsic variability in the simple IA network.

machine, connections between units are assumed to be symmetric: That is, the weight from unit *i* to unit *j* must be equal to the weight from unit *j* to unit *i*. This is already the case in the IA network we have been considering. Thus, we need not make any changes to the structure of the network. In particular the Boltzmann machine retains, and indeed requires, the interactive nature of processing inherent in the IA framework.

There are some differences between Boltzmann machines and IA nets. First, in the Boltzmann machine, units can take on only discrete activation values of 1 or 0. They do so according to the following formula:

$$p(a = 1) = \frac{e^{net_i/T}}{e^{net_i/T} + 1} \qquad (3)$$

The parameter T in this equation, called *Temperature*, is a scale factor that determines how gradually the probability will shift from 0 to 1 as the net input increases. At high temperature, probability shifts gradually with increasing net input. As T approaches 0, the transition becomes increasingly abrupt.

There is another difference between the Boltzmann machine and IA networks. In IA networks, processing is *synchronous*, in that each unit updates its activation at time t based on the activations of all units at time $t-1$. In the Boltzmann machine, processing is *asynchronous*, in the following sense: Units are chosen for updating one at a time, and as soon as a unit is updated its activation is used in all subsequent updates. Order of update is strictly random: At each time step, one unit is chosen for updating. First its net input is comp'ıted, and then the update equation given above is applied.

The net input to a unit is defined as before, with one slight change. In place of the decay toward resting level found in the IA update equation (Equation 1 on page 12), the Boltzmann machine makes use of bias terms for each unit. These can be set to negative values to keep units from coming on very often unless there is rather strong positive input from other units. Thus the net input to each unit is:

$$net_i = \sum_j w_{ij} a_j + ext_i + bias_i$$

The first thing we can observe about the Boltzmann machine is that the Boltzmann update equation (Equation 3) has a form similar to the Luce choice rule. The unit can be seen as choosing between two outputs (1 or 0). The excitatory inputs vote for the 1 response, and the inhibitory inputs vote for the response of 0.

When a Boltzmann machine is run at some temperature T>0, the network will eventually reach equilibrium (Hinton and Sejnowski, 1983). Often, equilibrium is reached by a process known as simulated annealing, in which the temperature is reduced in small steps from a high initial value to a low, final temperature. Annealing is, strictly speaking, not necessary for reaching equilibrium, though in practice it can take astronomical numbers of time steps to reach this state unless annealing is used.

At equilibrium, the activations of the units in the network are still subject to random fluctuation, but they fluctuate in a way that can be characterized elegantly. Imagine enumerating all the possible states of the network, such that each state represents a different assignment of the values 0 and 1 to the activations of the units. Then at equilibrium, the probability that the network is in a particular state x is given by:

$$p_x = \frac{e^{-G_x/T}}{\sum_k e^{-G_k/T}} \tag{4}$$

here G_k, called the *Goodness* of state k, is equal to:

$$G_k = \sum_{i<j} w_{ij} a_{ik} a_{jk} + \sum_i a_{ik}(bias_i + ext_i) \tag{5}$$

Here a_{ik}, a_{jk} represent the activations of units i and j in State k. Note that the first summation runs over all distinct pairs of units i,j only once.

The goodness of a state represents the extent to which the constraints that are represented by the weights and bias terms in the network, together with the external input to the network, are satisfied when the activations of the units in the network have the particular values that they have in the state. We can think of the positive weights as constraints indicating that both units connected by a weight should be on; and we can think of the negative weights as constraints indicating that at least one of the units connected by the weight should be off. Similarly, we can think of the bias and the external input to a particular unit as representing constraints indicating whether the unit should be on or off.

Consider now the simple network that we have been working with. Let us suppose that we specify a context that sends external input of 1 to one of the three context units /s/, /p/, or /t/, along with input to the target units. As before, we assume that the input to the target units consists of an input to each unit such that the sum of the two inputs is 1. We assume that we run the network until equilibrium is reached at some fixed temperature T. Then the state of the network is sampled, and the choice is made in accord with the state of the network at the moment of sampling: If at the moment of sampling the /r/ unit is on and the /l/ unit is not, /r/ is chosen; if /l/ is on and /r/ is not, /l/ is chosen. If neither unit is on or if both are on, a tie is declared and a choice is made at another, later random time sufficiently distant in time to be independent of the first. Under these conditions, the probability that response r is chosen is just the sum of the probabilities associated with being in each of the states in which r is on and l is off, divided by the sum of the probabilities of being in each of the states where either /r/ is on and /l/ is off, or /l/ is on and /r/ is off:

$$p(r) = \frac{\sum_{s_r} p_{s_r}}{\sum_{s_r} p_{s_r} + \sum_{s_l} p_{s_l}} \tag{6}$$

Here s_r ranges over states in which unit /r/ is on and unit /l/ is off, and s_l ranges over states in which /l/ is on and /r/ is off.

Let us consider some state s_r. Its probability is:

$$p_{s_r} = \frac{1}{Z} e^{G_{s_r}/T}$$

where Z just represents the denominator of Equation 4.

Now, the Goodness of such a state includes as one of the terms in the second summation in Equation 5 a term of the form $a_r(bias_r + ext_r)$, and another term of the form $a_l(bias_l + ext_l)$. Let us pull these terms out of the summation, and represent them as a_rR and a_lL; where R and L are ext_r+$bias_r$ and ext_l+$bias_l$ respectively. Then we can express the goodness of a state as the sum of three terms:

$$G_s = G'_s + a_rR + a_lL$$

where G'_s is just all the remaining terms in the goodness after a_rR and a_lL have been pulled out. Now, consider a state s_r in which the activation of the /r/ unit is 1 and the activation of the /l/ unit is O. For this state, the goodness is just $G'_{s_r} + R$. Using again the fact that the exponential of a sum is the product of the exponentials, we arrive at the expressions

$$p_{s_r} = \frac{1}{Z} e^{R/T} e^{G'_{s_r}/T}$$

and

$$p_{s_l} = e^{L/T} e^{G'_{s_l}/T}$$

Substituting into the expression for $p(r)$ given above, the $\frac{1}{Z}$ cancels out, and we find that

$$p(r) = \frac{e^{R/T} \sum_{s_r} e^{G'_{s_r}/T}}{e^{R/T} \sum_{s_r} e^{G'_{s_r}/T} + e^{L/T} \sum_{s_l} e^{G'_{s_l}/T}}$$

This expression is equivalent to the expression for $p(r)$ that we get from the Luce choice model. It states that the response strength associated with each alternative consists of a product of two terms, one associated with the input to the unit representing the alternative and one associated with the degree of contextual support for the alternative. This contextual support can be visualized in the following way. Suppose that we list all of the states of the network in which /r/ or /l/ but not both are on. For each r-state, there is a corresponding /l/ state that differs from the r-state only in the activations of the /r/ and /l/ units. Now consider G'_{s_l} and G'_{s_r} for each of these two states. The only differences between these partial goodnesses must involve connections between the /r/ and /l/ units and other units in the network, since these are the only two units whose activations differ between the two cases. For example, the positive connection between the unit for /r/ and the word unit for /tr/ would contribute to the partial Goodness of a state in which the /tr/ unit and the /r/ unit were on and the /l/ unit was off, but not to the partial goodness of another state in which the /tr/ unit is on but the /l/ unit is on and the /r/ unit is off.

To make this concrete, consider the following specific situation with respect to the network shown in Figure 5. The excitatory weights on the connections between levels are all 1, and the inhibitory connections within levels are all -1. The context units have biases of -.5 and external inputs of either 1 (for the unit that should be on in the context) or 0 (for all others). The target units have biases of -.5 and external input $.5+v$ to the unit for /r/ and $.5-v$ to the unit for /l/, $-.2<v<.2$. The word units have biases b which we will set to -1.9, and the Temperature, T, is set to .1. Under these conditions, most of the states of the network have negligible probabilities. The only states having non-negligible probabilities all have the correct context unit on and no other context unit on, as well as one of the two target units on but not both. If there is a word consistent with the context and the active target letter, the unit for that word may be on or it may be off, but no other word units will be on in either case. Specifically, if the context is /s/ the non-negligible states are:

(1) /s/ on and /r/ on with Goodness $.5+v$.

(2) /s/ on and /l/ on with Goodness $.5-v$.
(3) /s/ on and /l/ on and /sl/ on with Goodness $.5-v+k$.

Here $k=2+b$ which in this case is equal to .1. The strengths associated with these states are equal to

$$
\begin{aligned}
&(1)\ e^{(.5+v)/T} = e^{v/T}e^{.5/T}\\
&(2)\ e^{(.5-v)/T} = e^{-v/T}e^{.5/T}\\
&(3)\ e^{(.5-v+k)/T} = e^{-v/T}e^{(.5+k)/T}
\end{aligned}
$$

Plugging into Equation 6, $e^{.5/T}$ cancels out and we get

$$p(r)=\frac{e^{v/T}}{e^{v/T}+e^{-v/T}(1+e^{k/T})}$$

Here, then $e^{v/T}$ reflects the input support for /r/; $e^{-v/T}$ reflects the input support for /l/; and $1+e^{k/T}$ reflects the contextual support for /l/, which exceeds that for /r/ by the quantity $e^{k/T}$. This expression has the same form as the expression for response choice probabilities in the Luce choice model.

Choice Probabilities for Internal Units

The analysis given so far establishes a relation between the choice model and a Boltzmann machine for the case in which the units representing the alternatives we are interested in are direct recipients of external input. This is, not, of course a realistic assumption; we would in general suppose that the units representing response choices would be embedded deep inside a multi-layer processing system. It turns out that the correspondence extends to such cases, as long as the following conditions hold:

(a) The set of units in the network can be partitioned into three sets:

1. those that represent the alternatives among which a choice is being made.

2. those that represent the bottom-up input to the alternatives.

3. those that represent the context in which the input is being interpreted.

(b) There are no connections between units representing the input and those representing the context.

(c) Response choices are made by sampling states of the network at equilibrium until a state of the network is encountered in which the unit for one and only one of the alternatives is active.

The interactive activation model of letter perception has the characteristics required to meet conditions (a) and (b) (see Figure 9, from McClelland, 1985), and these conditions also hold approximately in the TRACE model.[5] Conditions (a) and (b) ensure that the Goodness of a state of the network contains no terms involving a product of the activation of a unit from the context and a unit from the input. Condition (c) further ensures that the Goodness of the states that contribute to choice probabilities can be partitioned into three parts, corresponding to the three parts of the network:

(1) G_{b_s}, the part of the Goodness due to the bias associated with the active alternative.

(2) G_{c_s}, the part of the Goodness due to all the terms involving context units. These terms include terms in which activations of context units appear singly as well as all terms in which such units appear in pairs with each other or with the active alternative.

(3) G_{i_s}, the part of the Goodness due to all the terms involving the input units. These terms include those in which activation of input units appear singly as well as all terms in which such units appear in pairs with each other or with the active alternative.

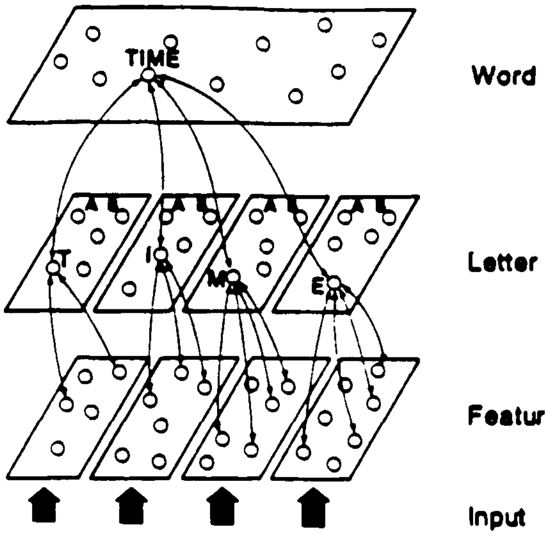

Figure 9. A sketch of the architecture of the interactive activation model of letter perception (McClelland and Rumelhart, 1981). This Figure is reprinted, with permission, from McClelland, J. L. Putting knowledge in its place: A framework for programming parallel distributed processing networks on the fly, *Cognitive Science*, 1985, p. 115.

[5] Actually there is a version of the TRACE model, TRACE I, in which they are clearly violated, but we will not be concerned with that version of the model here.

We wish to establish that Boltzmann machines that adhere to these assumptions exhibit two characteristics of systems that adhere to the Luce choice model: First, the probability of choosing alternative x is proportional to the strength of alternative x divided by the strengths of all of the other alternatives:

$$p(x) = \frac{S_x}{\sum\limits_k S_k}, \tag{7}$$

here k ranges over all of the alternatives, and where the strengths are assumed to be positive. Second, the strength of an alternative S_x can be written as the product of positive terms B_x, I_x, and C_x. The terms reflect the independent contributions of the bias in favor of alternative x, the input support for alternative x, and the contextual support for x:

$$S_x = B_x I_x C_x \tag{8}$$

The term B_x is assumed to be a characteristic of the alternative itself, independent of context and of stimulus input. I_x is assumed to vary with changes in the input, and C_x is assumed to vary with changes in the context; each is assumed to be independent of the other.

We wish to establish that a Boltzmann machine that conforms to (a)-(c) must also adhere to Equations 7 and 8 and exhibits independence of the terms B_x, C_x, and I_x. As a first step, we note that the probability of choosing alternative x is just the probability that the network is in a state associated with alternative x, divided by the sum of the probabilities that it is in a state associated with any of the alternatives. Using p_x to represent the probability of being in a state associated with alternative x, we have:

$$p(x) = \frac{p_x}{\sum\limits_k p_k} \tag{9}$$

The probability that the network is in a state associated with alternative x is just the sum of the probabilities associated with being in each of the particular states associated with alternative x.

Based on Equation 4, we can express this probability as:

$$p_x = \frac{1}{Z}\sum_i e^{G_{x_i}/T} \tag{10}$$

If we define the strength of alternative x to be the summation in the right-hand side of this equation:

$$S_x = \sum_i e^{G_{x_i}/T} , \tag{11}$$

then Equation 9 ..duces to equation 7. So all we have left to show is that S_x can be written as the product of the three independent terms B_x, I_x, and C_x.

Each term of the sum in Equation 11 can be represented as the product of the exponential of each of the three parts of the Goodness previously enumerated, so we have

$$S_x = \sum_i e^{G_{b_{x_i}}/T} e^{G_{c_{x_i}}/T} e^{G_{i_{x_i}}/T}$$

The first term, $e^{G_{b_{x_i}}/T}$ is constant in all elements of the summation, so we can pull it across the summation. This term is the required term B_x and is just equal to the exponential of the value of the bias on the unit that represents alternative x, divided by T, $e^{b_x/T}$. Pulling this out, we get

$$S_x = B_x \sum_i e^{G_{c_{x_i}}/T} e^{G_{i_{x_i}}/T} \tag{12}$$

The states of the network over which this summation runs can be laid out as an m by n matrix, where the m columns correspond to the m distinct states of the units in the context, and the n rows correspond to the n distinct states of the units in the input. These $m \times n$ states are all of the possible states in which x is on and the units representing the other alternatives are all off.

Now, given that conditions (a), (b), and (c) above all hold, the Goodnesses of the context states are independent of the input and of the input states. This is because no terms in the inp t show up in the context and no terms in the context show up in the input, as a result of the architecture of the net. So, $G_{c_{x_i}}$ is the same for all of the entries in a particular column of the matrix of states; by the same token, $G_{i_{x_i}}$ is the same for all the entries in a particular row. Index the columns by j and the rows by k, and designate the Goodness of the context for all the states in column j $G_{c_{x_j}}$ and the Goodness of the input for all of the states in row k $G_{i_{x_k}}$. The entry in cell jk of the matrix of terms in the expression for S_x is then just $e^{G_{c_{x_j}}/T} e^{G_{i_{x_k}}/T}$. The sum of all the terms in a column is just $e^{G_{c_{x_j}}/T} \sum_k e^{G_{i_{x_k}}/T}$. By summing these terms over columns, we can replace the summation in Equation 12 with $(\sum_j e^{G_{c_{x_j}}/T})(\sum_k e^{G_{i_{x_k}}/T})$, so the expression for S_x now becomes:

$$S_x = B_x(\sum_j e^{G_{c_{x_j}}/T})(\sum_k e^{G_{i_{x_k}}/T})$$

The first term we have already discussed. The second term (the summation over j) is the desired term C_x, reflecting only the context; and the third term, the summation over k, is the desired term I_x, reflecting only the input. All three terms are independent of the others, due to the architecture of the net.

Thus, it has been shown that response choices derived from equilibrium states of a Boltzmann machine with an architecture that conforms to *conditions* (*a*), (*b*), *and* (*c*) *above conform to the* relevant property of the Luce choice model.

Discussion. To summarize this section, we have found that the IA network shown in Figure 5 can implement the biasing effect of context that is described by classical models. We have also found that Boltzmann machines that follow the architecture of Figure 5, or even the more general architectural constraints given by conditions (a) and (b) above, also exhibit the characteristic properties of classical models.

It is not analytically obvious why the simple IA network of Figure 5 adheres to the classical models, when the standard processing assumptions of the IA model are used. This adherence seems to hold, but it is based on simulation. It remains to be seen whether this correspondence can be established analytically and whether there are any important boundary conditions on the circumstances under which it obtains.

In the absence of such results, we cannot simply leap to the conclusion that any reasonable size processing system such as the one embodied in the TRACE model will implement classical contextual influences on perception under the processing assumptions of the IA framework. We could move in the direction of ensuring this by replacing the dynamic assumptions of TRACE with those of the Boltzmann machine, while retaining TRACE's basic architecture. However, the equilibrium distributions of Boltzmann machines are themselves idealizations of real network behavior corresponding to statistical properties of ensembles of asymptotic states. It thus becomes important to be sure that in actual processing the right behavior is obtained. Here it is not clear that the Boltzmann machine assumptions are advantageous. Settling to equilibrium in Boltzmann machines typically requires a gradual reduction of Temperature from a high starting value, and this does not seem to be necessary with IA networks. Given this, and the prior history of success using the processing assumptions of the IA framework, I have chosen to continue simulations using the IA formulation.

Intrinsic Noise in TRACE

There are several differences between the situation in the TRACE model as described in McClelland and Elman (1986) and the situations we have examined above. First, the results thusfar ignore the fact that in domains such as speech, context typically precedes and/or follows the target stimulus. In visual presentation conditions, it is true, context and target are presented simultaneously. But in speech, the arrival of the stimulus -- context plus target -- is distributed over time. It is worth making sure that in this case, and in particular in the case where the context precedes the target, that the results already reported still obtain.

Second, the simple network in Figure 5 does not really have a word level; the top level might better be called a letter cluster level. In TRACE, knowledge of phonological regularities such as those captured by these letter clusters is distributed among the word units that happen to contain the relevant clusters. The simple network also ignores the presence of the feature level of processing altogether, and generally has an extremely simplified form compared to the much fuller, richer situation that must clearly obtain when real speech sounds are processed in context. While the original TRACE model is of course quite a bit simpler than any real perceptual process could be, it is nevertheless considerably richer and more complex than the tiny IA network that we have considered up to this point.

To check that these differences do not prevent TRACE from capturing the classical biasing effect of context, a simulation of the effects of intrinsic noise in the full TRACE model described in McClelland and Elman (1986) was undertaken. I first describe TRACE briefly then indicate that changes necessitated for the introduction of intrinsic variability.

The Structure of TRACE. The TRACE model consists of three levels, one for features, one for phonemes, and one for words. Units at each level are used to allow the model to represent what features, phonemes, and words may be present in each small time slice of a stream of spoken input. There is a separate unit for each feature in each time slice. Similarly, there is a separate unit for each phoneme in each time slice. Feature-level time slices are finer grained than phoneme level slices; each phoneme level slice extends over 3 feature-level slices. Words span sever^l phoneme-level slices, but for each word there is a unit for each possible slice in which the word could start. Input is presented (in the form of external inputs to feature units) a time slice at a time to successive banks of input units, so that it unfolds over time as it would if it were actually spoken or played back from a recording.

Bi-directional excitatory connections allow mutually consistent units to excite each other. Thus, the unit for /k/ in time slice *t* has mutual excitatory connections to units for words which contain a /k/ at time slice /t/. This includes the unit for *kup* starting at *t* and the unit for *stop* starting a t-1, for example. The /k/ unit also has mutually excitatory connections with units for the features of /k/ extending over several successive feature-level slices centered under time slice *t*.

Within each time slice at the feature level, units are further organized into dimensions. Within each feature dimension, units representing altemative values on the dimension in the same time slice are mutually inhibitory. At the phoneme level, units representing altemative phonemes in the same time slice are mutually inhibitory. At the word level, there is also mutual inhibition between word units proportional to the number of time slices of overlap between the words.

Simulation procedure. For the purposes of this simulation, syllabic contexts like those used by Massaro (in press) were employed. The contexts were /s_i/, /p_i/, and /t_i/. Here /i/ represents the vowel sound in the word *bee*. The input to the target phoneme had only three levels, one of which was /l/-like, one /r/-like, and one intermediate. These inputs were combined factorially to make nine syllabic stimuli which were used as the inputs to the simulation model. As just mentioned the feature description of each phoneme was spread out over several time slices, but the amount of spread was reduced so that successive phonemes did not overlap, to eliminate the contamination of the input to the detectors for the target phoneme by the features of the context phonemes. (The variable *fetspread* in the simulation program was set to 3 for each feature dimension. This parameter detemines how far each feature spreads on each side of its peak.)

On each simulation trial, the input is presented, one time slice at a time, as though it was a sequence of successive time slices of real auditory input. The 3-phoneme syllabic input pattems were spread over a total of 18 time slices, preceded and followed by 12 time slices of inputs representing silence. The peak of the target phoneme occurred at time slice 24.

As soon as the presentation of an input stream began, the interactive activation process was started. On every time slice, each unit's activation is updated, according to the IA update equation (Equation 1). However, just before a unit's activation is updated, its net input is perturbed by a random sample of normally-distributed noise with mean 0. The standard deviation of this noise was chosen to be sufficiently small (.02) so that the extemal input for the context phonemes virtually always gave rise to stronger activation of the correct context phonemes rather than any others in the appropriate time slices, thereby allowing the context to be effectively unambiguous. Input for the phonemes /r/ and /l/ were modified to make them more similar than they had been in the original model, so that they would be confusable in the presence of this small amount of noise.[6] At each time step in processing, running average activations of the units representing the target phonemes were calculated as described above, with the averaging-rate parameter λ set at 0.05. After a total of 90 time steps of processing, the running average activations of the units for /r/ and /l/ at time slice 24 were examined, and the alternative associated with the largest running average activation was chosen as the model's response.

Word units in the model were set up for the lexicon of 215 words used in McClelland and Elman (1986). The lexicon contained the words *sleep* and *sleet*, both of which were partially activated by the /s_i/ context; the word *tree*, which is strongly activated by the /t_i/ context; and several words beginning in /pr/ and /pl/. As it happened the lexicon contains more /pr/ words than /pl/ words, and the /pr/ words include the word *priest*, which was a better match to the /p_i/ context than any of the others since none of the others contained the vowel /i/.

The parameters used in the simulation were the same as those used in McClelland and Elman (1986) with the following changes: (1) the resting activation of word level units was reduced to -.1; (2) the phoneme-word and word-phoneme excitation parameters were set to 0.05 and 0.02; (3) the feature level decay was set to 0.02. These changes were necessary because with the original parameters the presence oi noise tended to cause the network to lock onto spurious words. All of the changes contribute to reducing this tendency.

500 trials were run in each of the nine conditions. The 4500 simulated trials (500 trials for each of 9 conditions) required approximately a week of computer time on a CONVEX C-1.

Figure 10 shows the z-transformed probability of reporting /r/ in each condition. The plotted points are fit with straight lines, choosing the spacing of the input conditions to promote a straight line fit. The fit to the z-transformed response probabilities is very close indeed, falls well within the standard error of each point, and accounts for 99.94% of the variance in the data. Thus it appears that even with all of the complexities of the TRACE model in place, the interactive activation process produces the pure biasing effect of context.

[6] The following represent the altered values of the external input on the ACUTENESS dimension for /r/, /l/, and the neutralized intermediate liquid:
0.0 0.0 0.0 0.0 .25 .50 .75 0.0 0.0
0.0 0.0 0.0 0.0 .75 .50 .25 0.0 0.0
0.0 0.0 0.0 0.0 .50 .50 .50 0.0 0.0
External inputs on all other dimensions were the same as in McClelland and Elman (1986).

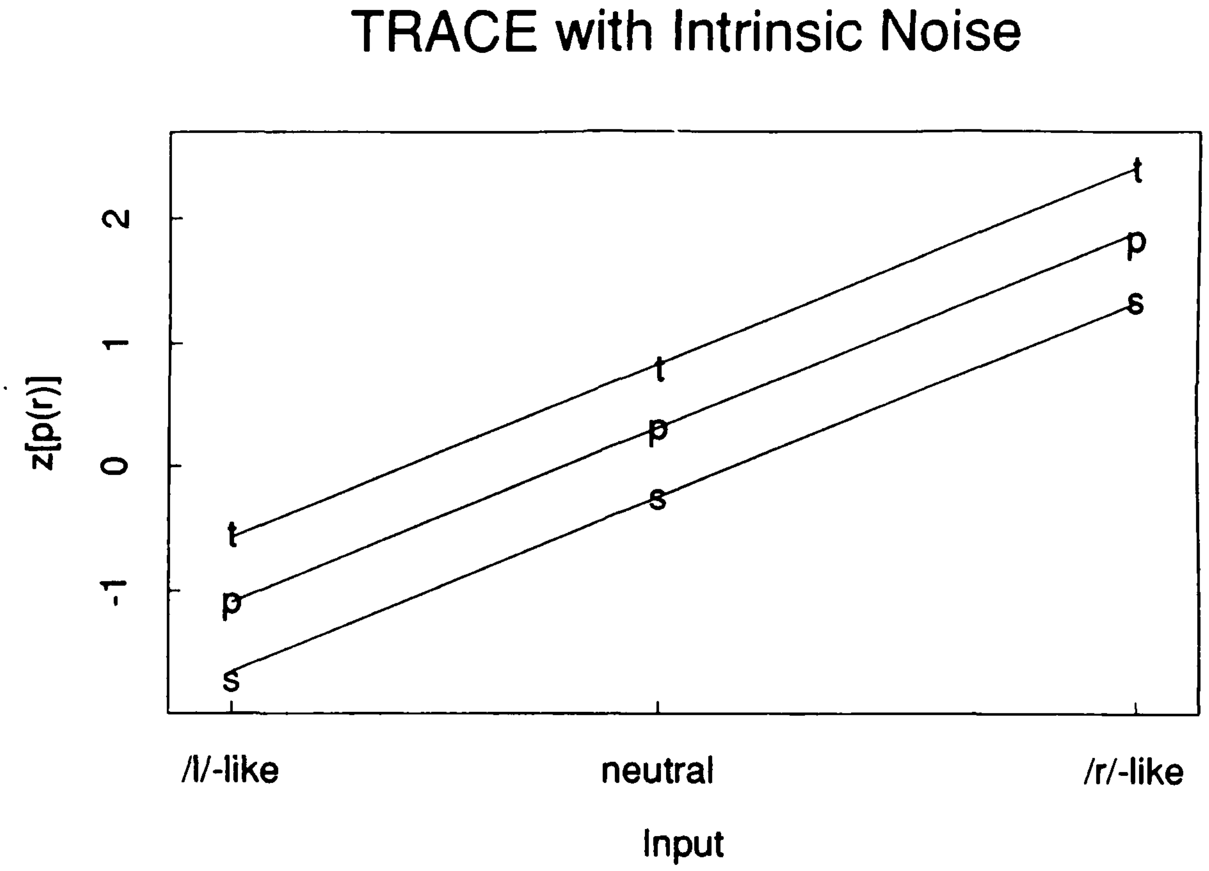

Figure 10. The joint effects of context and stimulus factors on the z-transformed probability of identifying the target letter as /r/, in the stochastic TRACE model.

General Discussion

The analyses and simulations described here indicate that interactive activation is not in fact incompatible with evidence that context exerts a biasing effect on perceptual identification responses. It is true that the assumptions of the IA framework, as originally formulated in McClelland and Rumelhart (1981), were flawed, but their flaw did not lie in their interactive character. This has been established by showing that compatibility with the classical biasing effect of context can be obtained, retaining the interactivity assumption that lies at the heart of the IA framework but replacing application of Luce's Choice rule to deterministic activations with a much simpler response selection rule applied to activations subject to variability. The source of variability may be either in the input to the network or in the processing mechanism itself. These findings have both practical and theoretical implications. I will consider these in turn.

Practical Implications

The findings have practical implications for efforts to simulate cognitive processes using the IA framework. While it shows that the framework is still viable, it means that deterministic simulations of the kind used in McClelland and Rumelhart (1981), Rumelhart and McClelland (1982) and McClelland and Elman (1986) cannot be expected to provide an accurate picture of all of the characteristics of the interactive activation process.

This implication is somewhat disconcerting, since it suggests that earlier results will have to be reassessed, and future results established, through computationally expensive stochastic simulations of the kind reported here rather than through deterministic simulations. While phenomenal improvements in computing technology have occurred since the original interactive-activation modeling work, stochastic simulations put a definite damper on the modeling process. What is to be done?

First, deterministic simulations need not be abandoned completely. They still provide a fairly clear picture of the time course of processing on a typical trial. They cannot be used reliably to indicate the probability of various outcomes, but practical experience suggests that they still provide a good guide to the median time it takes a unit inside a network to reach a particular level of activation. This result can be used to guide initial searches for parameter values that provide a pretty good fit to reaction time results. Noise can then be incorporated into the processing for further simulations in which reaction time distributions and error rates, as well as measures of central tendency can be considered.

Second, the study of very simple networks can allow the development of intuitions that can then be tested in larger models and in mathematical analyses. Simplification is important not only to avoid the computational expense associated with large stochastic networks, but also to reduce the problem to a sufficiently small scale so that it can be comprehended. I do not believe that the investigation reported here would have been successful had it not centered around the simple network of Figure 5.

Finally, more mathematical analysis will clearly be required. The analyses that I have been able to provide here (based on the idealization of networks as Boltzmann machines) applies only to equilibrium states. It would certainly be desirable to develop a characterization of the time course of processing in the same framework, and/or to get a firmer mathematical grip on networks that follow the processing assumptions of the original interactive activation model.

It seems unlikely that simulation will ever be replaced completely by mathematics. Mathematical analyses appear to be capable of establishing useful idealizations and boundary conditions, but do not necessarily provide the best guide to the practicality of ideas when they are actually put to the task of performing. Indeed in the end, it seems likely that stochastic simulations of small systems, deterministic simulations of full systems, and mathematical analyses will all need to be tested in the crucible of large-scale stochastic modeling.

Theoretical Implications

Now that we have established that interactive activation actually can produce the biasing effect of context that is described by classical models, we must ask, what implications does this have for theoretical conceptions of the process of perception? The first thing it suggests is that we do not need to view IA models as alternatives to classical accounts. Indeed it seems to me that both should coexist in psychological theory, each enriched and grounded by its relation to the other.

An alternative approach would be to take the fact that interactive activation implements the classical models as evidence that we need not be concerned with interactive activation. After all, it might be argued, psychologists need not be concerned with mere details of implementation. Why then should we not restrict our attention to the simple and elegant classical accounts?

The first point is that the classical models are not process models. As Luce points out, these models deal with asymptotic performance, not with the time course of information processing. In light of this, it seems natural to view interactive activation models as models of the time course of information processing, and to view classical models as succinct formal devices for characterizing the asymptotic results of information processing. The succinct formal characterization is very useful for many purposes, but a model of the underlying information processing has its uses too. For example, it holds out the hope of providing us with a means for understanding what is happening in cases where classical models do not provide a full account of the facts. There are findings in the literature which are not consistent with classical models. Among them is Samuel (1981) finding that d' is affected by context in certain phonemic restoration experiments. Indeed, classical accounts cannot explain the effect of word context on accuracy in forced-choice letter identification (Reicher, 1969). This effect was a large part of the impetus for the development of the interactive activation framework in the first place. In the terms of the present discussion, it seems reasonable to expect that a model that captures the time course of information processing would provide a better framework for accounting for the results of processing brief, masked stimuli since the stimulus conditions prevent processing from running to its asymptote. The interactive activation model of letter perception bears out this expectation (see McClelland and Rumelhart, 1981, and Rumelhart and McClelland, 1982 for discussion).

It is true that Massaro (in press) presents his fuzzy logical model as a process model. This model uses the same mathematics as Luce's choice model, but here the math is taken to characterize the outcome of processing in a feed-forward processing system containing separate stages of evaluation, integration, and decision.

The results reported here permit us to see that the fact that Luce's choice model fits asymptotic accuracy results does not favor the particular processing model Massaro proposes, relative, say, to an interactive activation account. Indeed we saw early on in this paper that a variety of assumptions about the role of context (Does it shift the criterion or add excitation to contextually appropriate detectors? Does variability influence the representations of input or only the choosing of responses?) are all compatible with classical context effects. The study of stochastic interactive activation and Boltzmann machines described in this paper shows further that these same classical effects are perfectly consistent with models in which context and target processing occur interactively. Thus the establishment of a correspondence between the mathematics of classical models and interactive activation indicates that there is no necessary implication from the adequacy of the Luce equation to the feed-forward processing assumptions that Massaro makes.

If classical context effects do not favor Massaro's model, they do not, in and of themselves, favor interactive activation modeis either. Obviously choosing between these two (and possibly others) will require additional research. There are some advantages to interactive activation models, though.

First, it must be noted that the originat motivation for the interactive activation model was specifically to address a considerable body of research in which the classical effect of context did not actually hold: A large number of experiments starting with Reicher (1969) demonstrated that word or pronounceable nonword context actually increases the accuracy of forced-choice identification of letters. This context effect is not simply a bias effect, and the interactive activation model of word perception provided a very good account of a number of experiments that obtained this effect.

Second, interactive activation models differ from Massaro's stage model in providing a mechanism for exploiting the mutual constraints each part of an input pattern impose on the interpretation of every other part. In most experiments, researchers have focused on the influence of context on the perception of some target item, and under these circumstances the fact that each part of the target-plus-context influences the perception of each other part is often lost from sight. But this mutual influence property can be observed in experiments in which the subject does not know which part of a displayed item wili be tested. Under these circumstances, increasing the duration of each letter influences the perception of every other letter (Rumelhart and McClelland, 1982, Experiment 6). The interactive activation model directly captures this, as each letter acts as context for the others; indeed, the assumption of interactivity was the basis for predicting the effect. It is difficult to see how this property could be captured by Massaro's stage model without allowing the results of the processes that he describes to be fed back -- i.e., without making the model interactive.

Third, the assumption of interactivity has led to another prediction that has been confirmed, this time in speech perception. The prediction is that compensation for coarticulatory influences of one phoneme on the acoustic realization of another could be triggered by context. The experiment relies on the fact that when we say a /t/ or /k/ following an /S/,[7] it is acoustically more /k/-like than it is when it follows a /s/, due to coarticulation of the two segments. Perceptually we compensate for this, so that a fixed segment that is perceived as neutral between /k/ and /t/ when preceded by silence will seem more /k/-like when spliced into a context where it is preceded by /s/ and more /t/-like when spliced into a context where it is preceded by /S/. Now, if we take a stimulus (which we will represent /X/) that is neutral between /s/ and /S/, and embed that sound in a context which favors /s/ (e.g., Christma_), an interactive account would predict that /s/ would become more active that /S/. This in turn would cause a following neutral /k/-/t/ stimulus to be perceived as /k/. Similarly a context favoring a /S/ interpretation of the /X/ (e. g., Spani_) should tend to cause perception of the following /k/-/t/ stimulus as /t/. This differential bias in the identification of ambiguous /k/-/t/ segments was confirmed in a series of experiments reported in Elman and McClelland (1988). Again, it appears that the perceptual results of contextual influences are being fed back into the processing system, exerting contextual influences themselves. In sum, the evidence is at the very least consistent with an interactive account of processing and may even tend to favor such an account over a strictly feed-forward processing system in some cases. Given this, it seems important to explore further the possibility that classical effects of context, when they do occur, may in fact be produced by an interactive activation system.

Conclusions

Massaro (in press) has helped to correct a deficiency of interactive activation models by pointing out that the IA framework, as originally formulated, produces incorrect simulations of classical contextual effects. Massaro's critique, combined with insights gained from simulations and relevant mathematical idealizations, have led to a correction of this flaw. The corrected model accurately accounts for the classical context effects.

The discovery that stochastic interactive activation actually does produce the classical effect of context is a step toward understanding the mechanisms of perception, but the step is a very small one. Further research is obviously necessary to establish whether stochastic interactive activation can still capture the findings encompassed by the interactive activation model of word perception and by the TRACE model of speech perception, and to see whether this framework can provide any insight into other situations in which classical models fail. If it should turn out that these extensions of stochastic interactive activation hold up in the face of the evidence, we will have in the interactive activation framework a model that provides a mechanistic

[7] We use /S/ to represent the sound associated with the *sh* in *ship*.

characterization of the processes that give rise to a wide range of empirical findings, extending well beyond the scope of what can be accounted for by the classical, asymptotic models. Of course the classical models will still characterize correctly the asymptotic outcome of processing in a wide range of cases. But they cannot provide a full characterization of the time course of processing. Further explorations within the interactive activation model will no doubt show that it has gaps in coverage and produces incorrect results at least in certain cases. Finding these gaps and discrepancies and exploring how they might be resolved should help us continue to move closer to an adequate model of the processes that allow context to influence perception.

References

Bagley, W. C. (1900). The apperception of the spoken sentence: A study in the psychology of language. *American Journal of Psychology, 12,* 80-130.

Elman, J. L., & McClelland, J. L. (1986). Exploiting the lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes.* Hillsdale, NJ: Erlbaum.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27,* 143-165.

Feldman, J. A. & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6.* 205-254.

Grossberg, S. (1978a). A theory of human memory: Self-organization and performance of sensory-motor codes, maps and plans. In R. Rosen & F. Snell (Eds.), *Progress in Theoret. Biol, Vol. 5* (pp. 233-374). New York: Academic Press.

Grossberg, S. (1978b). A theory of visual coding, memory, and development. In E. L. J. Leeuwenberg, & H. F. J. M. Buffart (Eds.), *Formal theories of visual perception.* New York: John Wiley and Sons.

Hinton, G. E., & Sejnowski, T. J. (1983). Analyzing cooperative computation. *Proceedings of the Fifth Annual Conference of the Cognitive Science Society.*

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in boltzman machines. In D. E. Rumelhart, J. L. McClelland & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: Bradford Books.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology: Vol I.* New York: Wiley

Massaro, D. W. (1975). *Experimental psychology and information processing.* Chicago: Rand McNally.

Massaro, D. W. (1979). Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 595-609.

Massaro, D. W. (in press). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology.*

Massaro, D. W., & Cohen, M. M. (1983). Phonological constraints in speech perception. *Perception & Psychophysics, 34,* 338-348.

McClelland, J. L. (1985). Putting knowledge in its place: A scheme for programming parallel processing structures on the fly. *Cognitive Science, 9,* 113-146.

McClelland, J. L. (1987). The case for interactionism in language processing. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading.* London: Erlbaum. 1-36.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception, Part I: An account of basic findings. *Psychological Review, 88,* 375-407.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises.* Boston, MA: MIT Press.

Miller, G., Heise, G., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41,* 329-335.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165-178.

Neisser, U. (1967). *Cognitive Psychology.* New York: Appleton-Century-Crofts.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172-191.

Pillsbury, W. B. (1897). A study of apperception. *American Journal of Psychology, 3* (3).

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81,* 274-280.

Rumelhart, David E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance VI.* Hillsdale, N.J.: Erlbaum.

Rumelhart, D. E. & McClelland J. L. (1981). Interactive processing through spreading activation. In C. Perfetti & A. Lesgold (Eds.), *Interactive processes in reading*. Hillsdale N.J.: Erlbaum.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception, Part II: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89*, 60-84.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474-494.